מבוא ללמידה עמוקה
תרגיל: 1
מגישים: יפעת חדד והדר שרביט
תאריך: נוב׳ 21

חלק מעשי

תהליך העיבוד המקדים

תהליך העיבוד התחיל ביצירת פיצ׳רים מתאימים לקלט. בהינתן רצף של 9 חומצות אמינו, יצרנו 180 פיצ׳רים בשיטה של one hot representation באופן הבא: לכל תו x שמייצג חומצה אמינית הנמצא באינדקס i ברצף התווים שמרכיבים את הפפטיד, יצרנו את הפיצ׳ר ix, שמשמעותו שבאינדקס i ישנה החומצה x. כך מרחב הפיצ׳רים שלנו הוא בגודל

$$|\{x : x \text{ is an amino acid}\}| \times |\{i : 0 \le i \le 8\}| = 20 \times 8 = 180$$

לדוגמא, הרצף $ABCDEFGHI$ ״הדליק״ רק את הפיצ׳רים $0A, 1B, 2C, \ldots, 8I$ (דהיינו, ערך מטריצת הפיצ׳רים במקומות אלה עבור הדגימה הנ״ל היא 1, וביתר 171 המקומות הערך הוא 0).

בסיום תהליך זה קיבלנו מטריצת פיצ׳רים $\mathcal{X} \in \mathbb{R}^{\#\ samples \times 180}$ (כאשר $\#\ samples \sim \underbrace{24{,}500}_{neg} + \underbrace{3000}_{pos}$)

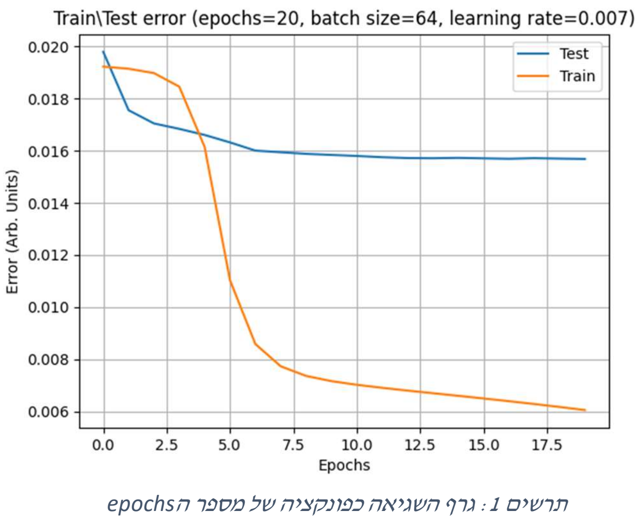
בחירת הארכיטקטורה

בהתאם לנאמר בהרצאה, תיעדפנו רשתות עמוקות על פני שטוחות, ומכאן המודל ההתחלתי אתו עבדנו היה מורכב מ-2 hidden layers. עקרון נוסף אותו בחנו היה שרשתות עם hidden layer שמרכיב יותר nodes מאשר שכבת הקלט נוטות לספק תוצאות טובות יותר. זאת בהתאם למה שנאמר בהרצאה – הרחבת ממד הקלט, היא שיטה נפוצה בקרב מספר לא קטן של ארכיטקטורות. עבור פונקציות המעבר – לאחר כל שכבה הופעלה פונקציית $relu$, אשר פעלה באופן גורף יותר טוב מפונקציית $sigmoid$. זאת פרט לפונקציה שלפני הפלט, אשר כן הייתה $sigmoid$, וזאת על מנת לנרמל את ערכי הרשת לכדי ערכים הסתברותיים בטווח $[0,1]$. גורם נוסף אותו בדקנו לאורך תהליך הלמידה הוא סוג פונקציית ה-loss איתה עבדנו. בתחילה השתמשנו ב-ℓ_2 סטנדרטי, אך לבסוף, משום שאופי הבעיה הוא של קלסיפיקציה עבור שתי מחלקות $\{pos, neg\}$, בחרנו בפונקציית שגיאה מסוג binary Cross Entropy.

בהקשר זה נציין שעל מנת להתמודד עם חוסר האיזון בין הדגימות ($24{,}500$ דגימות שליליו לעומת $3{,}000$ חיוביות), נתנו משקל גבוה יותר לדגימות החיוביות. בפרט, המשקל אותו בחרנו לתת היה אחיד בין הדגימות וערכו היה היחס בין הדגימות השליליות לחיוביות (פקטור של 8.2 לערך לטובת הדגימות החיוביות). נציין בנוסף שנעשה בתחילה שימוש במעברי $dropout$, כמו גם בשכבות $BatchNorm$, אם כי אלה לא השפיעו במיוחד על התוצאה הסופית.

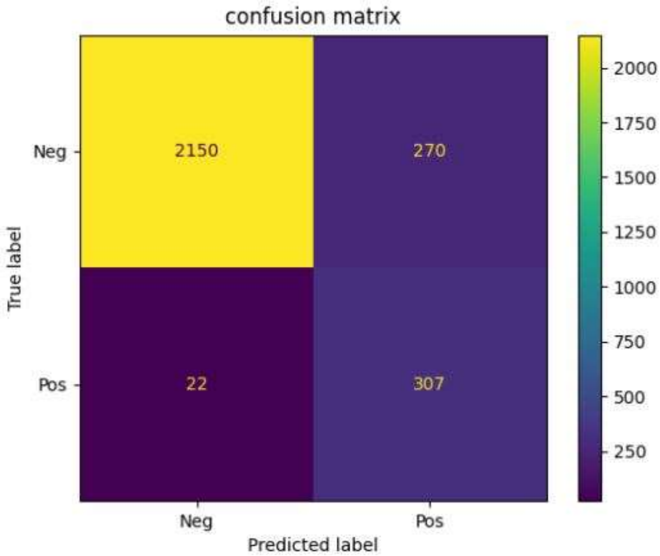
ניתוח תוצאות

על אף ניסיונות מגוונים, נראה כי שגיאת הטסט שאפה ל$0.0155\sim$ בקירוב ולא ירדה מעבר לכך. אנו סבורים כי הסיבה לכך נובעת בעיקר מאי יכולת הרשת להכליל בצורה מספיק טובה עבור דגימות חיוביות, וזאת למרות העובדה שהללו ממושקלים באופן משמעותי יותר ביחס לדגימות השליליות



תרשים 1: גרף השגיאה כפונקציה של מספר הepochs

חשוב להבין כי תרשים 1 לא מציב את כל התמונה, וניתוח מעמיק יותר לתוצאות ניתן לקבל מגרף ה confusion matrix ומחישוב יחסי ה-accuracy, הrecall וכדומה, כפי שנראה בפסקה הבאה :



תרשים 2: confusion matrix plot

מתרשים 2 ניתן להבין :

$$\text{True Positive} = 307, \text{False Negative} = 22$$

$$\text{True Negative} = 2150, \text{False Positive} = 270$$

כידוע, נוכל לרשום

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
$$\text{F}_1 = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}$$
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total}}$$

ולקבל מפורשות את הגדלים הבאים :

	precision	recall	f1-score	support
0.0	0.99	0.89	0.94	2420
1.0	0.53	0.93	0.68	329
accuracy			0.89	2749
macro avg	0.76	0.91	0.81	2749
weighted avg	0.94	0.89	0.91	2749

תרשים 3: classification report

נוכל לתת פירוש מילולי לתוצאותינו –

מדד *Precision* נותן את החלק היחסי עבורו צדקנו על דגימות חיוביות מתוך סך הדגימות החיוביות. במילים אחרות, זהו מדד ליכולתנו לתייג דגימות חיוביות כחיוביות מתוך כל הדגימות שברשותנו. במקרה הנידון היחס הוא

53%, כלומר יכולת התחזית שלנו עבור דגימה חיובית טובה מתחזית רנדומלית רק ב-3%. ניתן לשייך אחוז נמוך זה בעיקר לפער בין חלקיות הדגימות החיוביות והשליליות. עם זאת, יש לציין כי פתרון טריוויאלי המחזיר את התשובה $+1$ לכל קלט היה נותן ערך *precision* גבוה, וכמובן אין הדבר מעיד על איכות המודל, ובכל זאת – היינו רוצים לתייג נכונה את כל הדגימות, ולא רק חצי מהם.

כמו כן, מדד Recall נותן את החלק היחסי עבורו צדקנו על דגימות חיוביות מתוך סך הדגימות שתייגנו שהן חיוביות (בין אם צדקנו או טעינו). כאן התוצאות "טובות" יותר, עם 93%. המשמעות היא שברוב המקרים בהם זיהינו שפפטיד מיוחס לקורונה, אכן צדקנו. (אם כי שוב חשוב להדגיש שישנן דגימות חיוביות שכלל לא זיהינו)

נתבונן גם במדד ה accuracy, אשר נותן את אחוז הפעמים בהן צדקנו. במקרה שלנו הערך הוא 89%, כלומר כמעט ב-90% מהמקרים החזרנו תחזית נכונה. ואמנם מדובר באחוז גבוה, אך נזכור שרוב הדגימות הן שליליות ובאותה מידה רשת שמחזירה פתרון "טריוויאלי" של -1 הייתה מקבלת accuracy גבוה.

<u>היפר פרמטרים</u>

כפי שרשום בתרשים 1, ישנם מספר היפר-פרמטרים שנבחרו לשלב האימון. הראשון מביניהם היא מספר ה-epochs. את פרמטר זה בחרנו להיות המקסימלי כך ש-(1) לא נגיע למצב של overfit ו-(2) לא נגיע לunderfit. בהרצה של מספר פעמים נראה היה שהערך 20 היא המתאים ביותר. הפרמטר השני הוא ה-batch size – בחרנו אותו באופן דומה, כך שמצד אחד הוא קטן מדי מה שיגרור מיצוע על מעט מדי דוגמאות (דהיינו high variance) ומצד שני הוא לא גדול מדי מכדי למנוע רשת עם יכולת אקספרסיבית גבוהה (דהיינו להימנע מhigh bias).

הפרמטר האחרון הוא קצב הלמידה. ערכים שונים בטווח $[0.001,0.1]$ נוסו, אך די מהר התברר כי ערכים שגדולים מ-0.01 היו גדולים מדי, ואילו ערכים שקרובים ל-0.001 הובילו לתהליך למידה איטי מדי. לבסוף התברר הערך 0.007 כ-"מקום טוב באמצע".

<u>שאלה 6,7 - יצירת תחזית עבור חלבון Spike</u>

בהינתן רצף החלבון הנתון, ייצרנו מטריצת דגימות X וביצענו פרדיקציה בעזרת הרשת המאומנת. מתוך הפרדיקציה חילצנו את הפפטידים שזוהו עם הערכים המספריים הגבוהים ביותר והדפסנו אותם:

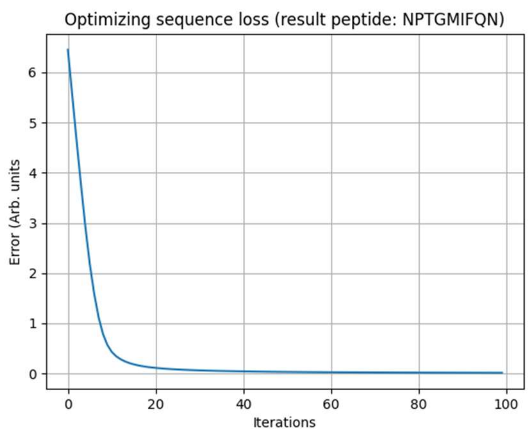
```
top five peptides: ['IPTNFTISV', 'LLFNKVTLA', 'FVFLVLLPL', 'PLVDLPIGI', 'FIAGLIAIV']
```

הקוד המתאים לחלק זה ממומש בפונקציה בשם

```
predict_peptide_from_spark
```

<u>שאלה 8 – *most detectable peptide*</u>

בחלק זה ביצענו איטרציות אימון על קלט רנדומי x שמייצג פפטיד יחיד. שגיאת האימון בשלב זה מתוארת בתרשים הבא



תרשים 4: *most detectable peptide loss error*

ופלט הפונקציה שניתן הוא (כפי שרשום בכותרת תרשים 4)

```
NPTGMIFQN
```

הקוד המתאים לחלק זה ממומש בפונקציה בשם

```
optimize_sequence
```

חלק תיאורטי

1. *פונקציות ליניאריות ואפיניות:*

יהיו $f(x)=Ax$ ו- $g(y)=By$ פונקציות לינארית כאשר $A\in\mathbb{R}^{m\times n}$, $B\in\mathbb{R}^{n\times k}$ ו-$y\in\mathbb{R}^k$, $x\in\mathbb{R}^n$.
פונקציית ההרכבה נתונה לפי

$$h := f\big(g(x)\big) = A\cdot Bx$$

נסמן $C:=A\cdot B$ כאשר $C\in\mathbb{R}^{m\times k}$ ונקבל ש-h היא פונקציה לינארית $h(x)=Cx$, כנדרש.

באותו אופן אם $f(x)=Ax+a$ ו- $g(y)=By+b$, כאשר $b\in\mathbb{R}^n, a\in\mathbb{R}^m$, הרכבה תיתן

$$h = f\big(g(x)\big) = A(By+b)+a = ABy+Ab+a$$

נסמן $C=A\cdot B$ ו-$c=Ab+a$ כאשר $C\in\mathbb{R}^{m\times k}$ ו-$c\in\mathbb{R}^m$ ונקבל ש-h היא פונקציה אפינית $h=Cx+c$ כנדרש

2. *אינפי של gradient descent:*

א. תנאי העצירה של הצעד $\theta^{n+1}=\varepsilon^n-\alpha\nabla f_{\theta^n}(x)$ יכול להיות למשל:

- היפר-פרמטר טריוויאלי שמגדיר מספר איטרציות (מספר epoch-ים):

For $i=1,2,\dots,T$:

$$\theta^{n+1}=\varepsilon^n-\alpha\nabla f_{\theta^n}(x)$$

- לולאה שעוצרת כאשר ה-גרדיאנט לא משתנה, או קטן מאיזשהו היפר-פרמטר מסוים ε:

While $\nabla f_{\theta^n}(x)>\varepsilon$:

$$\theta^{n+1}=\varepsilon^n-\alpha\nabla f_{\theta^n}(x)$$

- במקרים אידיאלים/ריאליזבילי הלולאה תעצור כשהגענו לנקודה סטציונרית
- ניתן לשלב את הטענות, למשל להגדיר מספר epoch-ים שבו נעצור בכדי להימנע מלולאות אינסופיות או ארוכות במיוחד (מה שיכול לקרות למשל כאשר הצעד α איננו אידיאלי)

ב. נשתמש בפיתוח טיילור לפונקציית ה-loss כדי לקבל תנאי לקבלת נקודה סטציונרית

$$f(x+dx)=f(x)+\nabla f(x)dx+dx^TH(x)dx+O(||dx||^3)$$

בנקודה סטציונרית מתקיים $\nabla f(x)=0$ ומכאן נשארנו עם

$$f(x+dx)-f(x)=dx^TH(x)dx$$

אם x היא נקודת מקסימום, כל הזזה dx תביא אותנו לנקודה עם ערך נמוך יותר, דהיינו $f(x+dx)-f(x)<0$. לעומת זאת, אם x היא נקודת מינימום, משיקולים דומים תקיים $f(x+dx)-f(x)>0$. כלומר בסה"כ התנאי הנדרש לסיווג הוא

$$dx^TH(x)dx=\sum_{i,j}\frac{\partial f}{\partial x_i\partial x_j}dx_idx_j\begin{array}{ll}>0: & maxima\\ <0: & minima\\ =0: & saddle\end{array}$$

3. *פונקציית הפסד למרחב מעגלי:*

לכל מרחק בין שני ערכים θ_1, θ_2 יש שני פירושים – הראשון הוא אורך הקשת הקצרה והשני הוא אורך הקשת הארוכה. לצורך חישוב מרחק מינימלי נרצה תמיד לבחור במרחק שמיוצג על ידי אורך הקשת הקצרה, ולשם כך נבחר במינימום בין המרחק עצמו (בזווית) לבין 360 פחות המרחק. פורמלית נרשום:
בהינתן מרחב דגימה $X = \{x \in \mathbb{R}: 0 \le x \le 360\}$ נגדיר פונקציית מרחק

$$\forall x_1, x_2 \in X\colon \operatorname{dist}(x_1, x_2) = \min(\{|x_2 - x_1|, 360 - |x_2 - x_1|\})$$

4. *נגזרות חלקיות*:

א. נסמן לשם נוחות $f_2 = x + y$ ו-$f_3 = 2x$

$$df_1 = \frac{\partial f_1}{\partial f_2} df_2 + \frac{\partial f_1}{\partial f_3} df_3 + \frac{\partial f_1}{\partial z} dz$$

ולכן בגזירה לפי x:

$$\frac{df_1}{dx} = \frac{\partial f_1}{\partial f_2} \underbrace{\frac{df_2}{dx}}_{=1} + \frac{\partial f_1}{\partial f_3} \underbrace{\frac{df_3}{dx}}_{=2} + \frac{\partial f_1}{\partial z} \underbrace{\frac{dz}{dx}}_{=0} = \frac{\partial f_1}{\partial f_2}(f_2) + 2\frac{\partial f_1}{\partial f_3}(f_3)$$

ב. נגזור לפי כלל שרשרת (כאשר סימנו $df/dx := f'$)

$$\tfrac{d}{dx} f_1\left(f_2(\dots f_n(x))\right) = \tfrac{df_1}{df_2}\left(f_2(\dots f_n(x))\right) \cdot \left(f_2(\dots f_n(x))\right)'$$

נחזור שוב על הפעולה עבור המוכפל הימני

$$\frac{d}{dx}\left(f_2(\dots f_n(x))\right) = \frac{df_2}{df_3}(f_3(\dots f_n(x))\left(f_3(\dots f_n(x))\right)'$$

כלומר

$$\frac{d}{dx} f_1\left(f_2(\dots f_n(x))\right) = \frac{df_1}{df_2}\left(f_2(\dots f_n(x))\right)\frac{df_2}{df_3}(f_3(\dots f_n(x))\left(f_3(\dots f_n(x))\right)'$$

נחזור על הפעולה באינדוקציה:

$$\frac{df_1}{dx} = \frac{df_1}{df_2}\left(f_2(\dots f_n(x))\right)\frac{df_2}{df_3}(f_3(\dots f_n(x))\left(f_3(\dots f_n(x))\right)' \cdot \dots$$
$$\cdot \frac{df_{n-1}(f_n(x))}{df_n}\frac{df_n(x)}{dx}$$

או בקיצור:

$$\frac{df_1}{dx} = \frac{df_1}{df_2} \cdot \frac{df_2}{df_3} \cdot \dots \cdot \frac{df_{n-1}}{df_n} \cdot \frac{df_n}{dx}(x)$$

ג. נגזור לפי כלל שרשרת, כאשר הגזירה יופיעו סכום הנגזרות החלקיות לפי כל פרמטר.
נשתמש בהגדרת הדיפרנציאל:

$$df_1 = \frac{\partial f_1}{\partial x} dx + \frac{\partial f_1}{\partial f_2} df_2 \to \frac{df_1}{dx} = \frac{\partial f_1}{\partial x}\frac{dx}{dx} + \frac{\partial f_1}{\partial f_2}\frac{df_2}{dx} = \frac{\partial f_1}{\partial x} + \frac{\partial f_1}{\partial f_2}\frac{df_2}{dx}$$

כלומר באופן מפורש:

$$\frac{d}{dx} f_1\left(x, f_2\left(x, f_3\left(\dots, f_{n-1}(x, f_n(x))\right)\right)\right) = \frac{\partial f_1}{\partial x} + \frac{\partial f_1}{\partial f_2}\frac{df_2}{dx}$$

נעשה את אותו תהליך עבור f_2:

$$\frac{df_2}{dx} = \frac{\partial f_2}{\partial x} + \frac{\partial f_2}{\partial f_3}\frac{df_3}{dx}$$

כלומר כעת הביטוי שבידינו הוא

$$\frac{df_1}{dx} = \frac{\partial f_1}{\partial x} + \frac{\partial f_1}{\partial f_2}\left(\frac{\partial f_2}{\partial x} + \frac{\partial f_2}{\partial f_3}\frac{df_3}{dx}\right) = \frac{\partial f_1}{\partial x} + \frac{\partial f_1}{\partial x} + \frac{\partial f_1}{\partial f_3}\frac{df_3}{dx} = 2\frac{\partial f_1}{\partial x} + \frac{\partial f_1}{\partial f_3}\frac{df_3}{dx}$$

נמשיך זאת באופן אינדוקטיבי, כאשר הנגזרת של f_i היא (לכל $i < n$)

$$\frac{df_i}{dx} = \frac{\partial f_i}{\partial x} + \frac{\partial f_i}{\partial f_{i+1}}\frac{df_{i+1}}{dx}$$

ולבסוף נישאר עם

$$\frac{df_1}{dx} = n\frac{\partial f_1}{dx} + \frac{\partial f_1}{\partial f_n}\frac{df_n}{dx}$$

ד. נסמן $\ell(x) = x + g\big(x + h(x)\big)$, וכעת השאלה היא

$$\frac{df\big(\ell(x)\big)}{dx} = \frac{df}{d\ell}\frac{d\ell}{dx}$$

את הנגזרת של ℓ נחשב מפורשות

$$\frac{d\ell}{dx} = 1 + \frac{d}{dx}g\big(x + h(x)\big)$$

נסמן באופן דומה $q(x) = x + h(x)$.

$$\frac{dg\big(q(x)\big)}{dx} = \frac{dg}{dq}\frac{dq}{dx}$$

את הנגזרת של q נחשב מפורשות

$$\frac{dq}{dx} = 1 + \frac{dh}{dx}$$

לבסוף נאחד את כל הביטויים:

$$\frac{df}{dx} = \frac{df}{d\ell}(\ell)\cdot\left(1 + \frac{dg}{dq}(q)\cdot\left(1 + \frac{dh}{dx}(x)\right)\right) = \frac{df}{d\ell}(\ell)\left(1 + \frac{dh}{dx}(x) + \frac{dg}{dq}(q) + \frac{dg}{dq}(q)\frac{dh}{dx}(x)\right)$$

5. <u>האנטרופיה היחסית היא גודל אי שלילי</u>

בהינתן התפלגויות P, Q מעל אותו מרחב הסתברות $\mathcal{X}$, מגדירים

$$D_{KL}(P||Q) := \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right)$$

נשתמש בקירוב הלינארי של $\ln(x)$, ובפרט בכך שלכל x מתקיים $\ln(x) \le x - 1$:

$$\begin{aligned}-D_{KL}(P||Q) &= -\sum_{x\in\mathcal{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right) = \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{q(x)}{p(x)}\right) \le \sum_{x\in\mathcal{X}} p(x)\left(\frac{q(x)}{p(x)} - 1\right)\\ &= \sum_{x\in\mathcal{X}} q(x) - p(x) = \sum_{x\in\mathcal{X}} q(x) - \sum_{x\in\mathcal{X}} p(x) = 1 - 1 = 0\\ &\Rightarrow D_{KL}(P||Q) \ge 0\end{aligned}$$

כאשר המעבר האחרון נובע מכך שסכימת ערך פונקציית הסתברות על כל המרחב נותנת תמיד 1.

כמו כן, קל לראות שכאשר $\forall x \in \mathcal{X}: p(x) = q(x)$ מתקיים $D_{KL}(P||Q) = 0$, שהרי $\log\left(\frac{p(x)}{q(x)}\right) = \log(1) = 0$

6. <u>האנטרופיה היחסית היא פונקציה קמורה</u>

בהינתן 2 נקודות$(p_1, q_1), (p_2, q_2)$ שמייצגות כל אחת זוג של פונק׳ הסתברות נראה שמתקיים

$$D(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2) \le \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2)$$

לשם ההוכחה נציג את הלמה הבאה:

$$a\log\left(\frac{a}{b}\right) \le \sum_i a_i\log\left(\frac{a_i}{b_i}\right)$$

כאשר סימנו $a = \sum_i a_i$, $b = \sum_i b_i$ ונדרש $\forall i\ a_i, b_i \ge 0$

הוכחת הלמה:

נסמן $f(x) = x\log x$ ונקבל

$$\begin{aligned}\sum_i a_i\log\left(\frac{a_i}{b_i}\right) &= \sum_i b_i f\left(\frac{a_i}{b_i}\right) = b\sum_i \frac{b_i}{b} f\left(\frac{a_i}{b_i}\right) \ge bf\left(\sum_i \frac{b_i}{b}\frac{a_i}{b_i}\right) = bf\left(\frac{1}{b}\sum_i a_i\right)\\ &= bf\left(\frac{a}{b}\right) = a\log\left(\frac{a}{b}\right)\end{aligned}$$

כעת נסמן

$$a = \underbrace{\lambda p_1(x)}_{a_1} + \underbrace{(1-\lambda)p_2(x)}_{a_2}, \qquad b = \underbrace{\lambda q_1(x)}_{b_1} + \underbrace{(1-\lambda)q_2(x)}_{b_2}$$

נשכתב את הביטוי של האנטרופיה היחסית במונחי a, b:

$$
\begin{aligned}
D(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2) &= D(a||b) \\
&= \sum_{x\in\mathcal{X}} a(x)\log\left(\frac{a(x)}{b(x)}\right) \overset{lemma}{\leq} \sum_i a_i \log\left(\frac{a_i}{b_i}\right) \\
&= \lambda \sum_{x\in\mathcal{X}} p_1(x)\log\left(\frac{p_1(x)}{q_1(x)}\right) + (1-\lambda)\sum_{x\in\mathcal{X}} p_2(x)\log\left(\frac{p_2(x)}{q_2(x)}\right) \\
&= \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2)
\end{aligned}
$$

7. <u>התיאוריה של ב*Cybenko* ו-*Hornik* עבור פונקציית $relu$</u>

התיאוריה של ב*Cybenko* ו-*Hornik* עוסקת בפונקציות חסומות, ואילו $relu$ איננה חסומה. עם זאת, הפרש של 2 פונקציות $relu$ כן ייתן לנו פונקציה רציפה וחסומה כפי שנדרש מהתיאוריה. כלומר אם נסתכל על זוגות איברים מתוך $f(x) = \left(\alpha_1\sigma(w_1x+b_1) + \alpha_2\sigma(w_2x+b_2)\right) + \cdots$, כל צירוף של זוג איברים כנל שהוא מהצורה $\mathrm{relu}(w_ix+b_i) - \mathrm{relu}(w_{i+1}x+b_{i+1})$ מהווה פונקציה חסומה ורציפה, ועל כן בסך הכל הדרישה של התיאוריה מתקיימת. ביתר פירוט, קיימת בחירה של w_i ו-b_i כך שהפרש ה-reluים היא פונקציה חסומה (למשל אם לכל זוג $w_i = w_{i+1}$ ו $b_i \neq b_{i+1}$, במצב זה נקבל פונקציה מונוטונית עולה חלש כך ש $\sigma(-\infty) = 0$ ו-$\sigma(\infty) = b_i$)

8. מבנה הרשת העמוקה עם התחשבות הסימן הוא של 4 זרועות במקום 3, כאשר הזרוע החדשה h_{new} תפעל בדומה לזרוע העליונה h_1, רק עבור ביטויים עם מקדם שלילי:

- בזרוע התחתונה h_3 נעביר קדימה את הקלט המוזז
- בזרוע האמצעית h_2 אנו נחשב את הפונקציה האפינית שרלוונטית לכל נוירון
- בזרוע העליונה h_1 אנחנו סוכמים את כל הפלטים של שכבות הביניים עבורם $\alpha_i > 0$ והסכום יהיה רק על גורמים חיוביים
- בזרוע החדשה h_{new} אנו סוכמים את כל הפלטים של שכבות הביניים עבורם $\alpha_i < 0$ והסכום יהיה רק של גורמים שליליים

בסך הכל הוספנו עוד $O(n)$ נוירונים, לכן בסך הכל נשארנו עם $O(n)$ נוירונים